NATIONAL COMPUTATIONAL INFRASTRUCTURE

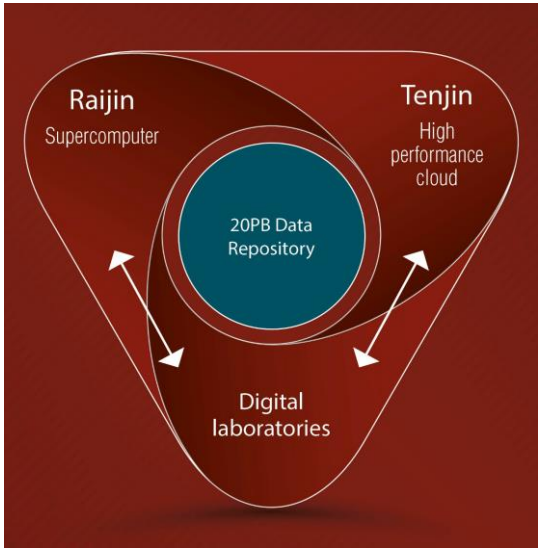
NCI's Research Data Services

Providing high-quality data to enable climate and weather science

- Claire Trenham

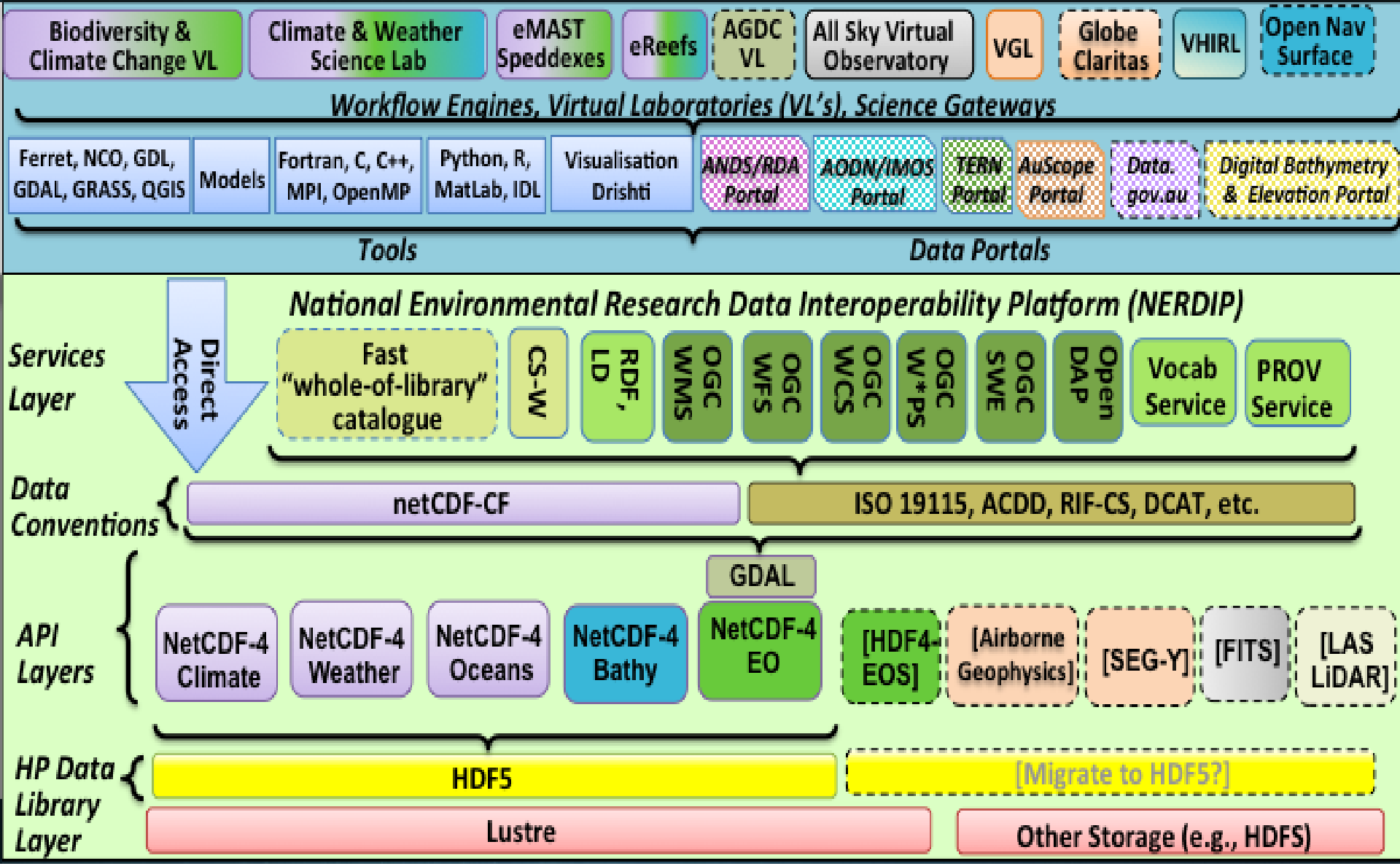
Kelsey Druken, Adam Steer, Ben Evans, Jon Smillie, Jingbo Wang

- NCI – National Computational Infrastructure
  - Highly integrated peak machine
    - Raijin: 1.2PFlops, >57k cores, Infiniband
  - data store
    - >30PB disk, ~10PB tape, 56Gb FDR Infiniband & 10GigE
  - research clouds
    - NeCTAR public cloud; Tenjin private cloud with Virtual Labs and access to 10+PB National Research Data Collection
  - Services
    - Academic consultants provide user support; scientific visualization; virtual laboratories; application optimization

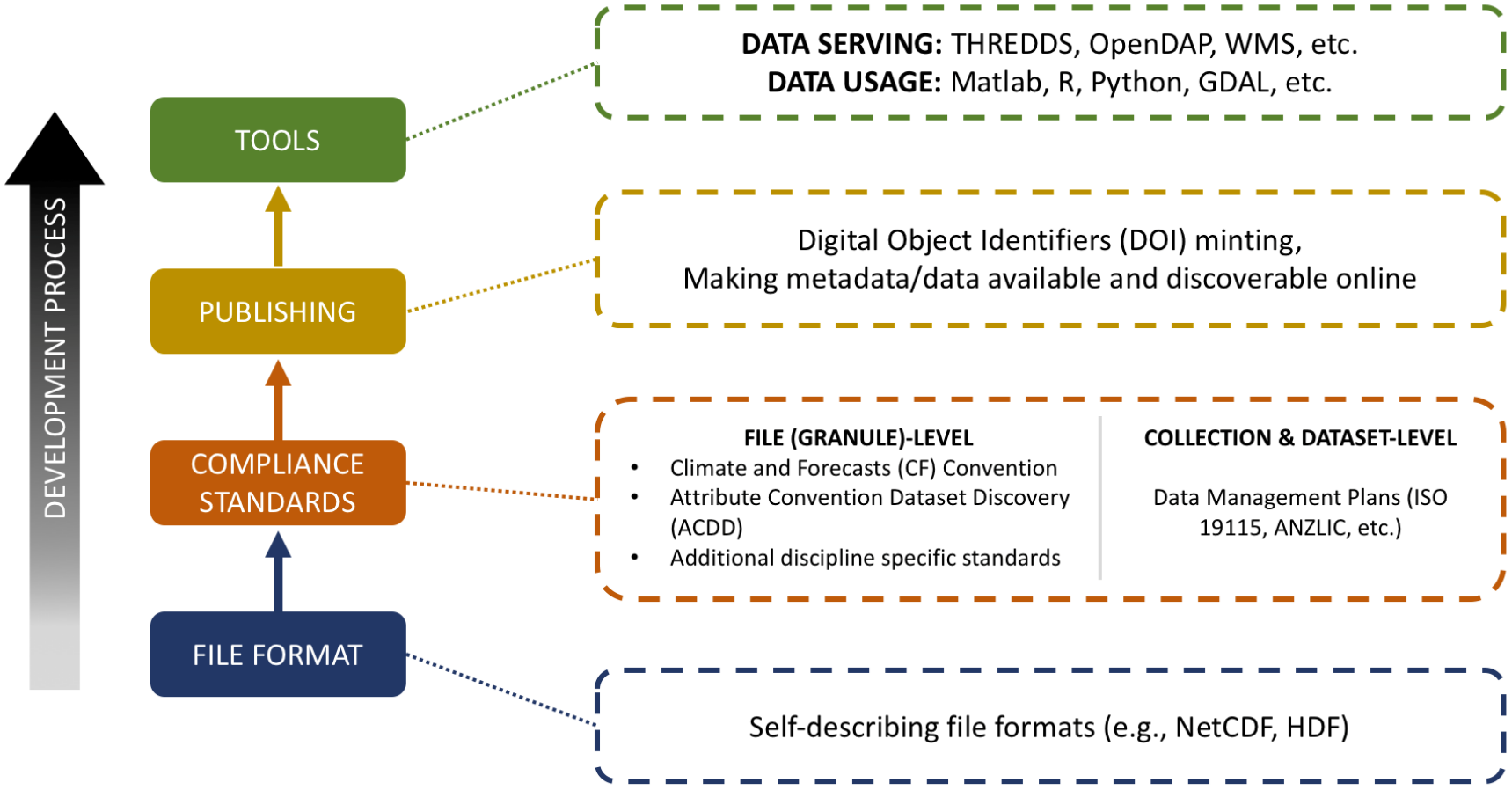


- RDS(I) funding provided to nodes around Australia for the storage of nationally significant data collections.
- NCI focus on the National Environmental Research Data Collection, comprising a range of fields including: climate, weather, Earth observations, ecology & land use, geophysics, geoscience, and astronomy; as well as data holding in social sciences, and bioinformatics.
- Over 10PB ingested and made available to community.
- Earth Systems Grid Federation primary node (climate models); Copernicus Hub for ESA data.

NCI National Environmental Research Data Interoperability Platform (NERDIP)



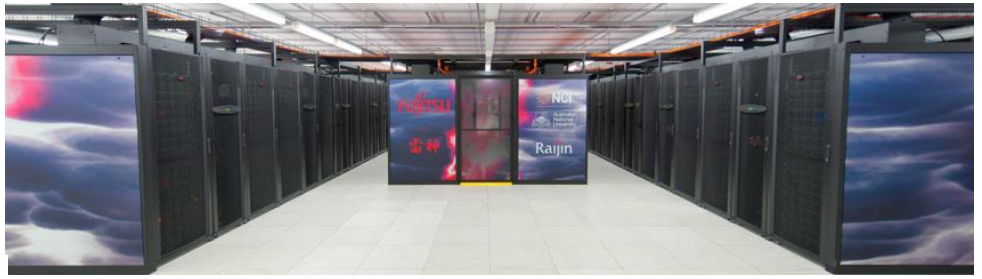
HPD Data Development Process



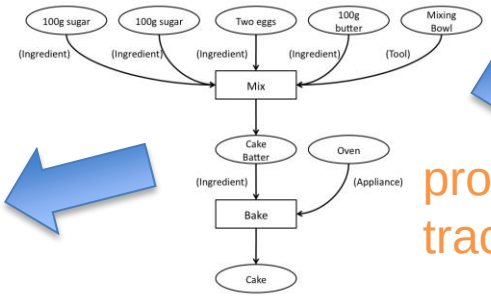
Data Ecosystem

data creation/
acquisition

data store



data processing pipeline



provenance
tracking

visualisation

publication
(data, code,
science results)

data access

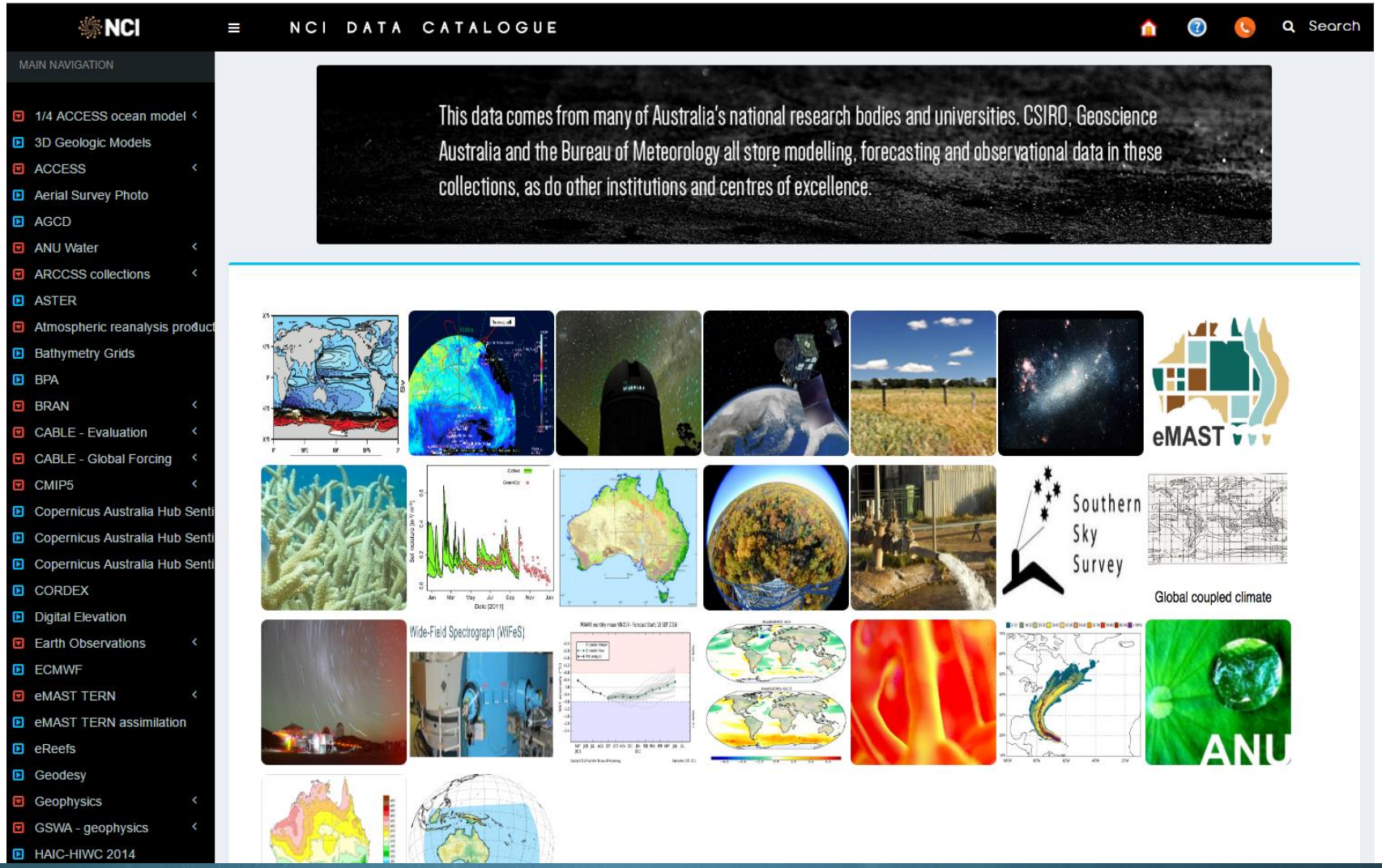
citation for ALL
your work

Data systems

- NCI have a multi-element system for metadata catalogues and data services
  - GeoNetwork, datacatalogue: Find metadata records
  - THREDDS Data Service: download or remotely access or view data
  - Geoserver (OGC web services), ESGF, ERDDAP, Hyrax, Rasdaman, and filesystem access
  - PROMS (provenance), PID service, RD Switchboard
  - DOI minting (citation)

Finding the (meta)data

https://datacatalogue.nci.org.au



Just get the bit of the data you need, not whole files!

- OPeNDAP – Network Data Access Protocol
  - Subset HDF5/netCDF files, only get bits of the data
- http://dap.nci.org.au THREDDS server
  - OPeNDAP/NetCDF Subset Server: subset, remote access
  - Other protocols supported by NCI
    - HTTP download
    - Open Geospatial Consortium Web Services (WMS, WCS, WPS, WFS…)
  - Underpins or provides data to a number of VLs
    - Virtual Geophysics Laboratory, WENFO, TERN, ANVGL, NEII, NationalMap, CWSLab, AusGIN, eReefs, BCCVL…

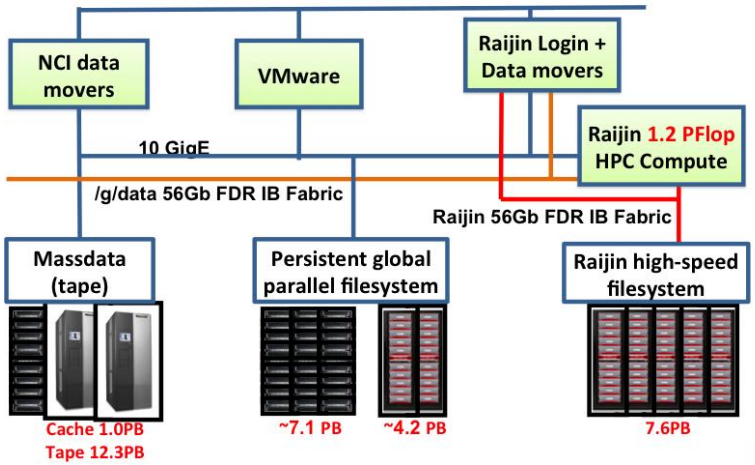
Accessing the data: OPeNDAP & NCSS

- OPeNDAP and NetCDF Subset Service allow subset selection and retrieval.
- Can access files directly from tools (Python etc.), use of Siphon package makes trawling directories much easier.
- Works with netCDF/HDF – standardizing formats is good!

Bring the scientists ***to*** the data!

Building The Platform for Earth System modelling & Analysis

Integrated HPC-HPD Environment



Data Services

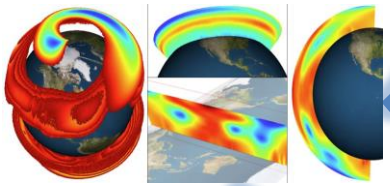
THREDDS

ESGF Earth System Grid Federation

OPeNDAP

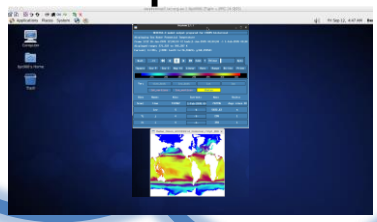
GeoServer

Server-side analysis and visualization



UV-CDAT

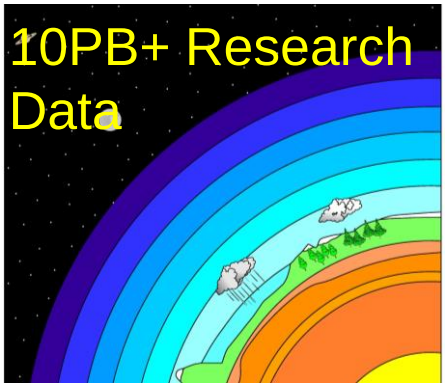
VDI: Cloud scale user desktops on data



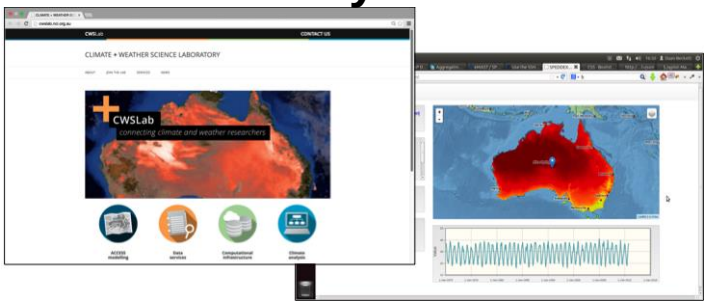
Puppet DevOps Framework

openstack

3000 Core Cloud

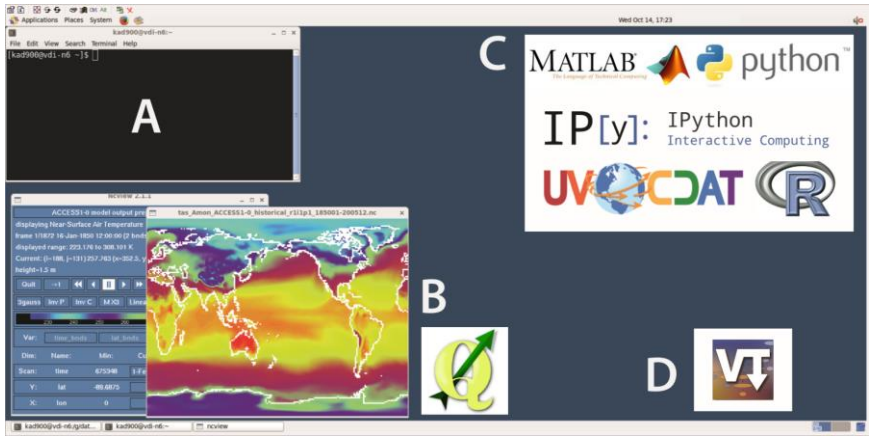


Web-time analytics software



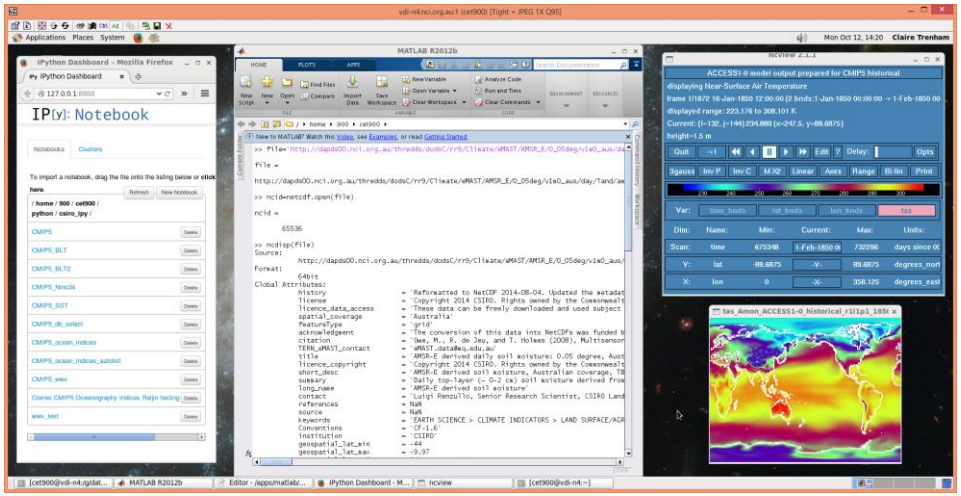
Bring the scientists to the data!

- Tools to support coding, data analysis & visualisation
- Virtual Desktop Infrastructure (VDI) to access, process & analyse data
  - CWSLab
  - AGDC



- Workflow tools allow community to implement analyses pipelines easily
- Many software tools available in this environment, integrated with the global Lustre filesystem and HPC infrastructure

Virtual Desktop Infrastructure



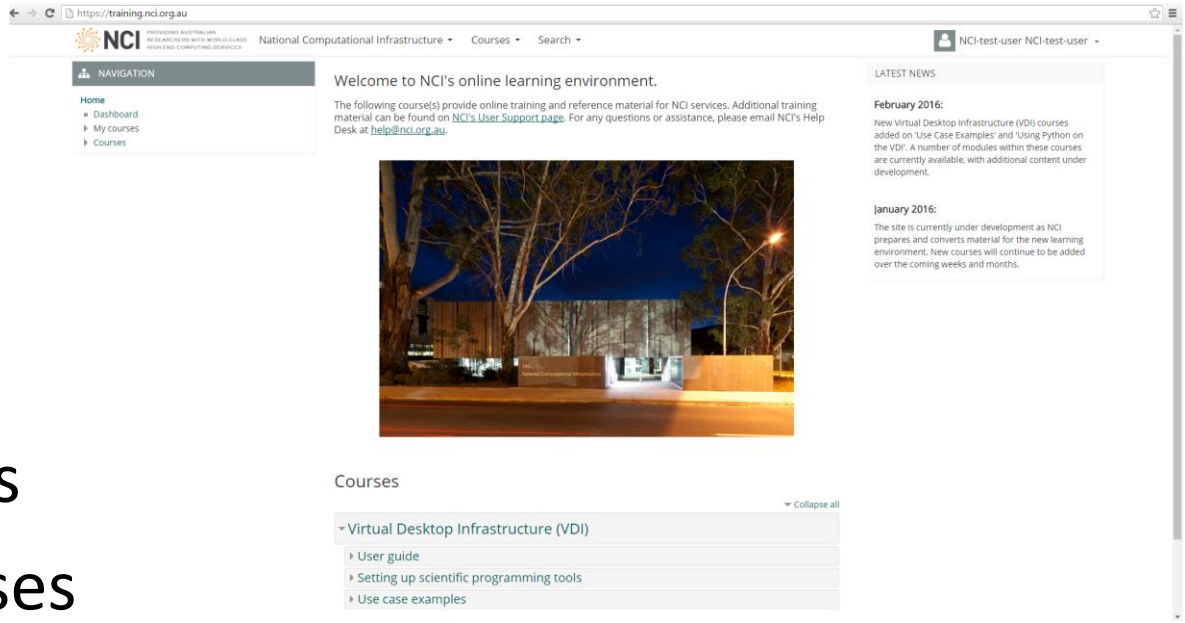
- Access granted per-project, data mounts as requested
- Software can be added as needed
  - Common science tools: python, Matlab, R, QGIS, compilers, libraries, …
  - pip + virtualenv to manage python packages on top of common python libs for climate science
- Desktops: 32GB RAM, 140GB local scratch, 8vCPUs, max session time 7 days
- User friendly, powerful, functional!

Risks of not centralizing the scientists to the data

- Data downloading and analysis by many users also has potential risks (apart from being too big for this to be feasible!)
  - Versioning of data used in analysis
  - Provenance tracking
  - Errata and Reporting
  - Documentation incorporated in file in case a file gets isolated?
- Bringing scientists to the data in Virtual Laboratories can mitigate these issues by ensuring everyone is working on the same data

Training users

- Unfamiliarity with such systems is a large hurdle in gaining access and confidence in using them.
- Established a Moodle LMS: https://training.nci.org.au
- Resources in
  - VDI use
  - Data finding
  - Data services
  - Python examples
  - Scientific use cases



Future directions?

- Web Processing Service to underpin on-demand computed products
- Birdhouse suite for specialised climate processing
- …?

- NCI THREDDS hosts Australian data contributing to CMIP5, CORDEX, GeoMIP, PMIP3, and soon CDR-MIP. ESGF node will publish same data – in progress!

- Thanks for listening!
- For further information and materials, please see the following links
  - https//nci.org.au
  - https://training.nci.org.au
  - https://esgf.nci.org.au
  - https://github.com/nci/Notebooks
  - https://nci.org.au/user-support/getting-help/
- Contact us: help@nci.org.au